DOCUMENT RESUME

ED 413 903 IR 056 733

AUTHOR Salaba, Athena

TITLE . Inter-Indexer Consistency in Bibliographic Record Access Points between OCLC and RLIN.

PUB DATE 1996-12-00

NOTE 46p.; Master's Research Paper, Kent State University.

PUB TYPE Dissertations/Theses (040) -- Reports - Research (143)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS Bibliographic Databases; *Bibliographic Records; Cataloging; Classification; *Indexes; Information Retrieval; Libraries; Library Catalogs; Online Catalogs; *Reliability; Search Strategies

IDENTIFIERS Access Control; *Inconsistency; OCLC; *Research Libraries Information Network

ABSTRACT

The purpose of this study was to examine the inter-indexer consistency in the assignment of selected access points between OCLC and RLIN (Research Libraries Information Network). Both databases are used by many libraries as bibliographic sources for copy cataloging. Access points were compared for differences in wording, MARC coding, and completeness. These access points included Library of Congress and Dewey classification numbers, main entry, uniform title, title statement and area of responsibility, title added entries, series, subject headings, and non-title added entries. Results showed a relatively high level of overall inconsistency (95%) among the access points analyzed. Largest numbers of differences were found in the subject heading fields (40.7%), the title statement and area of responsibility (18.8%), and series fields (13.6%). Findings are compared to the results of previous error studies. (Contains 17 references.)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

IR

ED 413 903

INTER-INDEXER CONSISTENCY IN BIBLIOGRAPHIC RECORD ACCESS POINTS BETWEEN OCLC AND RLIN

A Master's Research Paper submitted to the
Kent State University School of Library and Information Science
in partial fulfillment of the requirements
for the degree Master of Library Science

by

Athena Salaba

December, 1996

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
R. Du Mont
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

IR056733

BEST COPY AVAILABLE

ATHENA SALABA
"INTER-INDEXER CONSISTENCY
IN BIBLIOGRAPHIC RECORD
ACCESS POINTS BETWEEN
OCLC AND RLIN"
ADVISOR: DR. CARLYLE
DECEMBER 1996

ABSTRACT

One hundred and six matching bibliographic records from OCLC and RLIN were compared for inter-indexer consistency in the assignment of access points. Access points were compared for differences in wording, MARC coding, and completeness. Access points included Library of Congress and Dewey classification numbers, main entry, uniform title, title statement and area of responsibility, title added entries, series, subject headings and non-title added entries.

A high level of overall inconsistency (95%) was found among the pairs studied. The largest numbers of differences were found in the subject heading fields (40.7%), in the title statement and area of responsibility (18.8%), and series fields (13.6%).

The findings were compared to the results of previous error studies and conclusion were drawn wherever it was possible.

Master's Research Paper by

Athena Salaba

B.A., Kent State University, 1995

M.L.S., Kent State University, 1996

Approved by

Adviser [signature] **Date** 12-3-96

Table of Contents

List of Tables	iv
Introduction	1
Literature Review	3
Methodology	12
Data Analysis	20
Conclusions	31
References	36
Appendix	39

List of Tables

1a. LCCN Differences 20

1b. LCCN Consistency 21

2a. DDC Differences 21

2b. DDC Consistency 22

3. Choice of Main Entry Differences 22

4a. Main Entry Differences 22

4b. Main Entry Consistency 23

5a. Title Statement and Area of Responsibility Differences 23

5b. Title Field Consistency 24

6a. Title Added Entry Differences 24

6b. Title Added Entry Consistency 24

7a. Series Differences 25

7b. Series Consistency 25

8a. Subject Heading Differences 26

8b. Subject Heading Consistency 26

9a. Non-title Added Entry Differences 27

9b. Non-title Added Entry Consistency 27

10. MARC Differences 28

11a. Summary of Access Point Consistency 29

11b. Summary of Access Point Differences 30

CHAPTER I

INTRODUCTION

The rapid growth of new technologies enables us to automate almost all of the essential functions of a library. Today most libraries use an automated system for the organization and management of library materials. Users are able to search more than one library for the information they need. This leads us to the necessity of quality control of the access points. Even though most libraries use the Anglo-American Cataloguing Rules (AACR2) and the Library of Congress Subject Headings (LCSH), and follow in general the Library of Congress cataloging practices, we see differences among indexers in the assignment of access points. Studies of interindexer consistency show the degree of variability in access point assignment.

Even though most libraries use an automated system, not all libraries use the same automated system. Therefore, library users have to learn the different strategies and commands that each OPAC requires when searching for a specific item or topic. This can be very confusing, especially to the researchers who will probably have to use more than one library and more than one type of library (e.g. academic, special, public). Once the users learn the specifics of an OPAC they should be able to retrieve the material they are searching for by entering the same terms in the different OPACs, especially among libraries that follow the same cataloging principles.

The purpose of this study is to examine the inter-indexer consistency in the assignment of selected access points between OCLC and RLIN libraries. Both databases

are used by many libraries as bibliographic sources for copy cataloging. Therefore, not only libraries that originally catalog items and input records but also a large number of member libraries share these bibliographic records and copy them to their local databases. This study will discover the degree to which there is consistency in assigning access points and which access points show the greater variability.

CHAPTER II.

LITERATURE REVIEW

Inter-indexer consistency is a measure of the degree of agreement between two indexers in the assignment of codes or terms that represent a document. Earlier studies have shown that there is a positive correlation between inter-indexer consistency and information retrieval effectiveness (Chan 1984, Slamecka and Jacoby 1965). Therefore, studies of inter-indexer consistency are essential for higher quality control of access points and consequently for better services provided to library users and information seekers.

Furnas, Gomez, Landauer, and Dumais conducted four studies in which people would have to name several objects in such a way that others would be able to identify them. Their findings show that there is a great diversity in the way people described common objects. The results of these studies have very valuable implications for the design of retrieval systems. The lack of agreement in terms used between the designer and the user will lead to failure of information retrieval (Furnas and others 1982).

Inter-indexer consistency is commonly measured by the number of "exact matches" of index terms between indexers. Leonard lists six conditions that have to be considered before determining an "exact match." These are: 1) single and multiple word terms, 2) synonyms, 3) word endings, 4) different spelling of terms, 5) punctuation, and 6) abbreviations (Leonard 1977).

According to Markey (1984) inter-indexer consistency improves when indexers use controlled vocabularies to assign index terms. The better the familiarity an indexer has

with the structure of controlled vocabularies and the more experienced in indexing he/she is, the greater the inter-indexer consistency is. When inter-indexer consistency between experienced and inexperienced indexers was compared, the consistency scores varied only a little between the two groups. Markey, in her study, tried to test the findings of earlier studies of inter-indexer consistency when indexing print material to the inter-indexer consistency when indexing visual material. The difference between the two types of material is that the indexer of visual material has to determine the symbolic meaning first and then to transform it into subject headings. Her findings show even lower scores of inter-indexer consistency when indexing visual material than the scores of indexing print material. These low scores could be attributed to "the selection of artworks from a particular medium, chronological period, geographical area where lack of certain qualities could cause difficulty to indexers" (Markey 1984, 174). The use of uncontrolled vocabulary and indexers with no subject specialty could have attributed to the low inter-indexer consistency scores also.

Slamecka and Jacoby (1965) measured inter-indexer consistency by the number of terms used by each indexer for each document, and the percentage of matching terms assigned by experienced and inexperienced indexers. In their research of inter-indexer consistency they studied the effect of indexing tools. Among their findings are: 1) indexers assign fewer terms in average but have a higher inter-indexer consistency when using indexing aids, 2) the use of alphabetical vocabularies and classification schedules produces the highest average consistency between indexers, 3) indexing from titles,

abstracts or defined portions of a document produces higher consistency than indexing from the entire document.

In a similar study, Tell and Atonenergi (1969) compared the inter-indexer consistency when indexing from titles, abstracts, and from full text. For this study an authority list of index terms or keywords and the UDC classification system were used as indexing tools. Tell and Atonenergi found that "there is no significant difference in reliability when using just titles or just abstracts" but there was a "significant drop in reliability when using full text for indexing" (Tell and Atonenergi 1969, 287). They also found that indexing consistency is significantly lower when using keyword indexing than when using a hierarchical indexing system, UDC in this case.

Herr (1970), in her study of database access a training tool for indexers and a method for increasing indexing consistency, states that the goal of indexing is to assign index terms to a new document that reflect its relationship to the other documents of the collection. Database access, "the availability of all the indexing for a data base," allowed decentralized indexers to communicate with each other and new indexers to be trained with minimal contact with the experienced indexers. The study showed that there was no significant effect of the use of the database on the depth of indexing. Indexer consistency between the participants in this study and the reference indexer (experienced indexer used as a reference by new indexers) increased when the database was used as indexing tool.

MacMillan and Welt indicate that the human factor exists "even in similarly-educated, equally-trained indexers working in the same field" (MacMillan and Welt 1961, 89). In their analysis of the subject headings and sentence wording they noticed that the

differences in indexing were more obvious than the similarities. With the increase of the number of index terms per documents an increase in the inconsistency of index terms was observed.

Chan (1989) studied inter-indexer consistency in subject cataloging between LC and non-LC members of OCLC, when both are using the same tools and follow the same principles. Pairs of records for the same item were used in this study of LC subject headings. An average of 1.9 headings per record was assigned by non-LC libraries and 2.14 headings per record were assigned by LC. In other words LC records received an average of 12.63% more headings that the non-LC records. Only 15% of the records were perfect matches, 80% had some headings that matched completely or partially, and 5% of the records were totally unmatched. Of the 80% or the imperfectly matched records 57% had one or more, but not all, completely matched headings and 43% had only partially matched headings. Main headings varied in capitalization, punctuation, spelling, entry element, and wording. While variation in capitalization usually does not affect retrieval, variation in punctuation and spelling may have an affect on retrieval. Variation in entry element and wording definitely have an affect on retrieval. These findings suggest that even with the use of the same indexing tools and following the same indexing policies, total consistency is very difficult. The norm appears to be partial consistency among indexers.

A number of error studies are available which are very similar to inter-indexer consistency studies. A few error studies were focused on the study of access points.

These studies were very helpful in devising categories for types of access points to study in this present paper.

Romero and Romero (1992) studied the number of errors found in the MARC records in the catalog of the Library at the University of Illinois at Urbana-Champaign. The "categories of errors considered were errors in description, in headings, in classification, in subject headings, in capitalization and punctuation, in encoding the MARC format, including omission of required data" (Romero and Romero 1992, 51). The area of headings, which included main entry (1XX) and added entry (7XX) fields, accounted for 7.3% of all errors in all areas. Classification (DDC) accounted for 7.6% of all errors. Errors in the area of subject headings accounted for 7.9% of the total number of errors. The largest number of errors, 40.6%, was in the area of encoding of MARC format. The area of capitalization accounted for .6% of total errors. Finally, 15.4% of the total errors were accounted for by punctuation errors.

In a similar study, Intner used 215 matched pairs from OCLC and RLIN "for entries for books published in 1983 or later for which there was no LC record or other 'authoritative' cataloging and for which the member-contributed record was full-level cataloging" (Intner 1989, 38). Spelling, punctuation, capitalization, application of AACR2, application of LC rule interpretations, and MARC coding errors were counted in this study. The results show a 50.3% of total errors in the OCLC records and a 49.7% in the RLIN records, not a significant difference in accuracy or completeness between the two databases according to Intner. Of the total number of errors 37.1% was accounted for by application of AACR2 and LCRIs errors, 27.1% by punctuation errors, 16.7% by

MARC encoding errors, 8.4% by capitalization errors, 8.1% by spelling errors, 1.4% by call number errors, and 1.2% by subject errors (Intner 1989).

All of the above studies focus on the measurement of inter-indexing consistency or cataloging errors. Another focus of inter-indexer consistency research is the identification of the factors influencing inter-indexer consistency.

Chen and others (1994), in their study of managing inconsistency of information when using a Computer-Assisted Research System (CARS), observed two types of inconsistencies: subject-area-related and classification-scheme-related. Subject-area consistency was determined by the completeness and accuracy of uncontrolled keywords assigned to the text; classification scheme consistency was determined by the indexers' choice of folders and proper application of controlled vocabulary. Consistency was affected by the indexer's knowledge of the subject area and the classification scheme. Inconsistencies were due to subject area problems, classification scheme problems, system problems, and collaboration problems. Subject area problems were related to the uncertainty about keywording principles, such as use of abbreviations, reflection of the whole document, assignment of specific descriptors, etc.; choosing uncontrolled terms as keywords; and to the difficulties in interpreting the text. Classification scheme problems were related to the uncertainty of folder (category) contents, uncertainty about the folder assignment principles, and the uncertainty about controlled vocabulary assignments. There were two types of system problems: a) the use of inappropriate search options and b) the incidental system errors, such as incorrect typing. Collaboration problems were the

uncertainty about ongoing research topics and the insufficient communications among researchers.

Blair also suggests that the variation among indexers in the assignment of index terms to a document is a "function of the character of the subject area with which the documents to be indexed deal and the levels of training and experience of the indexers involved" (Blair 1986, 229). Another function that influences inter-indexer consistency is whether or not a controlled indexing vocabulary is used by the indexers. Blair states that it is unlikely for an indexer to know the entire set of candidate index terms for one document. In order to determine the candidate set, the same document would have to be indexed at least by several indexers. Even then, if additional indexers were to index the same document the candidate set would increase.

Another set of factors influencing inter-indexer consistency was studied by Tarr and Borko (1974). These factors are: 1) the recognition of indexable concepts; 2) the verbal paraphrasing of the concept, 3) the selection of an appropriate entry term, and 4) the translation into a standard term. In their study, Tarr and Borko used two groups of indexers to assign verbal labels (uncontrolled phrases) to documents and/or to translate the verbal labels into standard index term from the ERIC thesaurus. The first group of indexers had professional training and/or experience. The second group of indexers had no professional experience but had varying degrees of training. This study showed that there was no variance in inter-indexer consistency between indexer pairs; that is their conclusions are: 1) variations in inter-indexer consistency do not correlate with the characteristics of each document, 2) the use of a single indexing procedure leads to the

increase of indexer consistency, and 3) the use of controlled vocabulary improves indexer consistency only when indexing is done in two distinct but interrelated steps which are forming verbal labels first and then index terms by using the ERIC Thesaurus .

Finally, Zunde and Dexter (1969) group the factors influencing indexing consistency into three categories: 1) pragmatic, 2) semantic, and 3) environmental. Under the pragmatic category fall factors related to dispositions, capabilities, psychological conditions, background, etc. of each indexer. Pragmatic factors include educational background, maturity, personality, aptitude for indexing, motivation in work, fatigue, etc. Semantic factors are the author's vocabulary, readability of the text, subject area, specificity of discourse, indexing method, indexing aids, etc. Environmental factors include noise in the working area, climatic conditions, lighting conditions, time constraints, type of technical equipment, etc. In their study, Zunde and Dexter investigated the readability of the document being indexed, a semantic factor, and the effects of temperature in the working area, an environmental factor. The findings of the study of readability of a document show that there is some correlation between readability and quality of indexing. Indexers were more consistent when indexing material at the level at which they were the typical audience. The findings of the effects of temperature on indexer consistency show that the difference in average quality was not significant. There was more variation in quality from one document to another, than there was in the "average performance of an indexer under varying conditions" (Zunde and Dexter 1969, 321). Therefore they conclude that neither of the two factors influence significantly the indexing consistency.

All of the above studies show that there is a degree of inter-indexer inconsistency when assigning index terms, class numbers, or subject headings to documents. The present research studies the number of differences in assigning selected access points to books. Based on previous research it is expected that this study will find a varying number of differences among indexers for the different access point studied here. The results of this study will add significant new information to the field of inter-indexer consistency and will assist the research on information retrieval effectiveness.

CHAPTER III.

METHODOLOGY

This paper examines the inter-indexer consistency between OCLC and RLIN, using a methodology similar to the studies conducted by Romero and Romero (1992), Intner (1989), and Chan (1989). Specifically, it examines whether or not there are differences in the assignment of access points to bibliographic records by catalogers from libraries that are members of OCLC or RLIN. One hundred and six matched pairs of records for the same items, one from OCLC, one from RLIN, were analyzed. Access points including classification numbers, main entry, title statement and responsibility, title added entries, series, Library of Congress subject headings, and added entries are examined to determine whether there are differences in the presence or absence of a field, in the number of fields assigned for each type of access points, and in the content of each field.

Since it is impossible to analyze all records of both databases, a sample of bibliographic records was selected. In order to ensure that a variety of records were sampled, a systematic sampling method was used, consisting of selecting the first and last subject headings of every 25th page of the 18th edition of the Library of Congress Subject Headings. Each subject heading selected was searched first in RLIN because it allows easier subject retrieval. Because of the structure of the RLIN database, subject retrieval of all formats in one search is not possible. Therefore only books are included in this study. The sample includes the first record of each that satisfied the following parameters:

1. No records existed in the database from the Library of Congress or any other U.S. national agency. Many libraries copy or base their cataloging on LC records when these are available; therefore these records were not considered originally cataloged by the catalogers of the member libraries.

2. The records include Library of Congress subject headings. Because of their wide use, LC subject headings are most studied and therefore more consistency among catalogers is expected in the assignment of LC subject headings. This makes the study of inter-indexer consistency of LC subject headings a good contrast to other access points.

3. The records were entered in the databases after 1982. At the end of 1981 LC and ARL members started using AACR2. Because of the differences between AACR2 and AACR in the rules regarding the assignment of access points records following AACR2 are selected.

4. The original items are in English. Accurate cataloging of material in foreign languages depends on cataloger's knowledge of foreign languages. To avoid accounting for differences caused by the variability in the level of language knowledge, records of items in English are selected and analyzed in this study.

5. There is no OCLC number on the RLIN record. An OCLC number indicates that the record was copied form OCLC and therefore both databases have identical records for the specific item.

All the records that were selected from the RLIN database were then searched in OCLC by title. Exact matches that were non-LC records or created by a different agent than the RLIN records were selected from the OCLC database. OCLC records that represent microfilm or microform versions of books selected from RLIN were excluded. The two records for the same item, one from RLIN and one from OCLC formed a matched pair. The use of this method generated a sample of 106 pairs of matching records.

When all matched pairs were collected each pair was analyzed for differences on the following fields:

Library of Congress (LC) and Dewey (DDC) classification numbers (050, 055, 082, 090, 092)

Main entry (1XX)

Uniform title (130, 240)

Title and statement of responsibility (245)

Title added entries (246, 740)

Series (4XX, 8XX)

LC subject headings (6XX)

Non-title added entries (7XX)

Differences analyzed in all of these fields included word/number order, presence or absence of words, and MARC tagging differences. Spelling, punctuation and

capitalization errors were not counted. Each field was also analyzed for specific differences.

Libraries are not required to use the same classification system so inter-indexer consistency differences are expected in this category. Therefore, only pairs with LC or DDC numbers in both records were compared for classification number differences. LC classification numbers were compared by looking at class differences (the first letter part of LCCN, e.g. QA), class number differences (including the numerical part and the Cutter number for a topic), Cutter number differences (Cutter for main entry), and date differences. DDC numbers were compared by looking at class differences (the numeric part assigned from the DDC Schedules), Cutter differences (Cutter for main entry), and date differences. Difference in the use of 050 or 055 or 090 fields for LC number and in the use of 082 or 092 fields for DDC numbers was not counted as difference. In the following example of a pair of LC classification numbers zero class differences, one class number difference, one Cutter number difference and zero differences in date of MARC coding were found:

Record 1: 090 BX5129 $b .E39 1994

Record 2: 050 4 BX1765.2 $b .E383x 1994

Differences in the choice of main entry were counted and reported separately from title (245) or 1XX differences. When one record had a 1XX field and the other had a title main entry this counted as one choice of main entry but was also counted as one difference

in the 1XX field (because it was present in only one record of the pair studied). For example:

Record 1: 245 00 Karnataka Darshana

Record 2: 100 1 Hukkerikar, R. S.
245 10 Karnataka darshana

This pair had one difference in the choice of main entry but also one difference in the presence of 1XX field.

In addition to MARC differences 1XX fields were compared by looking at heading differences. In the following example only one heading difference was found:

Record 1: 100 1 Wislocki, George Bernays, $d 1892-1956.

Record 2: 100 1 Wislocki, George Bernays, $d 1892-

The access points of uniform title (130, 240) were compared by looking at title statement differences, and presence/absence differences.

The title and statement of responsibility areas (245) were compared to find whether there were differences in the title proper, subtitle or the area of responsibility.

Example:

Record 1: 245 10 Report on sound trials of the sirens at the south-east and north customs.

Record 2: 245 10 Report on sound trials of the sirens at the south-east and north-east Shantung promontories / $c published by order of the Inspector General Of Customs.

The above pair had one difference in the title proper, one difference in the area of responsibility and one MARC coding difference.

When comparing title added entries, the varying form of title field (246), and title added entry field (740) were treated as the same field because the usage of field 246 has changed recently. Because 740 field used to be used where now a 246 field is used, if one record traced an added title in a 246 field and the other record had a 740 field but the title statement was consistent then no differences were counted. Title added entries (246, 740) were compared by looking at presence/absence differences and title statement differences. The following example had one presence/absence difference:

Record 1: 246 30 Magic of the marque

Record 2: 740 01 Magic of the Marque
740 01 M G

The series fields were compared to find differences between the two records of each pair in the use of 4XX field only or a 490 field first and then an 8XX field (field treatment) and differences in series tracing practice, series statement, numbering, and presence/absence of each series statement. When two series were traced in one record and only one of them was traced on the other this counted as a series statement difference and also as a presence difference. In the following example one difference in the series statement and one difference in the field treatment were found:

Record 1: 440 0 Occasional papers of the Institute for Antiquity and Christianity ; $v no. 10.

Record 2: 490 1 Occasional papers of the Institute for Antiquity and Christianity ; $v no. 10.
830 0 Occasional papers (Institute for Antiquity and Christianity) ; $v no. 10.

LC subject headings were further analyzed for differences in the main heading, subdivisions (only when main headings were identical), and number of subject headings assigned. When one record had three subject headings and the other record had only two and only one heading was present and consistent in both records, this counted as a one difference in the number of headings assigned and three main heading differences. If the main heading was consistent in both records but the subheadings were different this was counted as a subheading difference only. For example, the following pair had two main heading differences and four subheading differences.

Record 1:
650 0 Celts.
650 0 Nationalism $z Ireland.
650 0 Nationalism $z Wales.
650 0 Nationalism $z Scotland.

Record 2:
650 0 Pan-Celticism.
650 0 Celts.
650 0 Civilization, Celtic.
650 0 Nationalism $z Great Britain.

Non-title added entries were compared for differences in main heading, subheading, and presence/absence of a field. For example, when:

Record 1: 700 10 Wiese, A. F.

Record 2: 700 10 Frans, Robert.
700 10 Wiese, Allen F.

this counted as one presence/absence difference and one main heading difference.

For each pair a difference worksheet (see Appendix 1) was used and differences in each field were counted. After all pairs were analyzed, the results were tabulated. Results show the number of differences within a field, the percentage of difference within

the field, and the percentage of difference of the total number of differences of all matched pairs.

CHAPTER IV

DATA ANALYSIS

Selected access points were compared in 106 matching pairs of OCLC and RLIN records. Data was analyzed by counting the number of differences in each access point between the pairs and by calculating simple percentages for each type of difference within each field. The percentage of difference for each type of access point out of the total number of differences was also calculated. Two tables presenting the data for each specific type of access point are given: first, one table for differences within the access point and second, one table for the level of consistency in the specific access point. MARC differences combined across the types of access points are presented in a separate table following the data analysis of the specific access points. Finally, differences among all of the access points are summarized.

Consistency in the assignment of classification numbers was analyzed first.

Table 1a: LCCN differences

LCCN (050, 055, 090)	# of differences	% of field differences	% of total differences
Class differences	5	17.86	1.23
Call number differences	8	28.57	1.98
Cutter number differences	9	32.14	2.22
Date differences	6	21.43	1.49
MARC	0	0	0
Total	28	100	6.9

A total number of 28 differences, 6.9% of the total number of all differences, were found among the 14 pairs of records with LC classification numbers. These include 5 differences in class number (17.86% of the differences within the field and 1.23% of the total number of all differences), 8 differences in call number, which represent 28.57% of the differences within the LCCN fields and 1.98 % of the total number of all differences, 9 differences in Cutter numbers (32.14% of

LCCN differences and 2.22 % of all differences), and 6 date differences which account for 21.43 % of LCCN differences and 1.49 % of the total number of all differences. There were no MARC coding differences.

# of pairs compared	14	13% of 106
# of pairs with at least 1 difference	12	86% of 14
# of pairs with >1 diff.	8	57% of 14
# of pairs consistent	2	14% of 14

Table 1b: LCCN consistency

Out of the 14 pairs (13% of 106) compared for LC numbers, 12 (86% of 14) had at least one difference (Table 1b). Eight of these 12 pairs (57%) had more than one difference, and only two pairs (14% of 14) were consistent.

DDC (082, 092)	# of differences	% of field differences	% of total differences
Class differences	1	50	.25
Cutter differences	1	50	.25
Date differences	0	0	0
MARC	0	0	0
Total	2	100	.5

Table 2a: DDC differences

Since there were only three matching records with DDC numbers present in both of them, the numbers reported were very small (Table 2a). There were only two differences in the assignment of DDC numbers, which represent .5% of the total number of all differences. Of the two, one difference occurred in the class number and the other in the Cutter number, each representing 50% of the differences within the field and .25% of total number of all differences. There were no date or MARC coding differences in this category.

# of pairs compared	3	2.8% of 106
# of pairs with at least 1 difference	1	33% of 3
# of pairs with >1 diff.	1	33% of 3
# of pairs consistent	2	67% of 3

Table 2b: DDC consistency

Among the three pairs (2.8% of 106) compared for DDC number differences one pair (33% of 3) had more than one difference and two pairs (67% of 3) were consistent (Table 2b).

Next, main entry fields were compared. First, all 106 pairs were compared to see if the choice of main entry was consistent. The comparison of the choice of main entry (1XX field versus 245 field) showed that seven pairs, 6.6% of the total number of pairs studied, had a main entry field difference, which was 1.7% of the total number of all differences (Table 3).

Main Entry	# of differences	% of 106 pairs	% of total differences
Total	7	6.6	1.7

Table 3: Choice of main entry differences

A total number of 16 differences occurred within the 1XX field, accounting for 3.7% of the total number of all differences (Table 4a).

1XX	# of differences	% field differences	% of total differences
Heading	13	80	2.97
MARC	3	20	.73
Total	16	100	3.7

Table 4a: Main entry differences

Of the 16 differences, 13 were heading differences, accounting for 80% of the differences in the 1XX fields and 2.97% of the total number of all differences. There were 3 differences in MARC coding, accounting for 20% of the total number of differences in 1XXs and .73% of the total number of all differences.

BEST COPY AVAILABLE

# of pairs compared	80	75% of 106
# of pairs with at least 1 difference	16	20% of 80
# of pairs with >1 diff.	0	0
# of pairs consistent	64	80% of 80

Table 4b: Main entry consistency

Eighty pairs (75% of 106) were compared for differences in the 1XX fields (Table 4b). Sixteen pairs (20% of 80) had at least one difference and no pair had more than one difference in the main entry field. Sixty-four pairs (80 % of 80) contained consistent 1XX fields.

A total number of 75 differences were found in the 245 field (title) among the 106 matching pairs, accounting for 18.8% of the total number of all differences (Table 5a). Eight of these differences occurred in the title proper, which were 10.95% of the total 245 differences and 1.98% of the total number of all differences. There were 16 differences in the subtitle, 21.92% of field differences and 3.96% of all differences. The statement of responsibility (respons.) had 34 differences accounting for 46.58% of field differences and 8.42% of all differences. Finally, 17 differences occurred in the MARC coding, 20.55% of field differences and 3.71% of all differences.

Title (245)	# of differences	% of field differences	% of total differences
Title proper	8	10.95	1.98
Subtitle	16	21.92	3.96
Respons.	34	46.58	8.42
MARC	17	20.55	3.71
Total	75	100	18.8

Table 5a: Title statement and area of responsibility differences

BEST COPY AVAILABLE

# of pairs compared	106	100% of 106
# of pairs with at least 1 difference	45	42 % of 106
# of pairs with >1 diff.	22	21 % of 106
# of pairs consistent	61	58% of 106

Table 5b: Title field consistency

The title statement and area of responsibility (field 245) are mandatory, therefore all 106 pairs were compared for differences in this field (Table 5b). Forty-five pairs (42% of 106) had at least one difference in the title statement or area of responsibility with 22 pairs (21% of 106) having more than one difference. Out of the 106 pairs 61 (58%) were consistent.

Title added entries (fields 246 and 740) were analyzed as one category (Table 6a). A total number of 11 differences (2.7% of all differences) were found in this category including 10 presence/absence differences, which accounted for 91% of title added entry differences and 2.5% of all differences, and 1 MARC coding difference, 9% of title added entry differences and .2% of the total number of all differences. There were no title statement differences.

Title added entries	# of differences	% of field differences	% of total differences
Presence/ absence	10	91	2.5
Title statement	0	0	0
MARC	1	9	.2
Total	11	100	2.7

Table 6a: Title added entry differences

# of pairs compared	14	13% of 106
# of pairs with at least 1 difference	11	79% of 14
# of pairs with >1 diff.	1	7% of 14
# of pairs consistent	3	21% of 14

Table 6b: Title added entry consistency

Fourteen pairs (13% of 106) were compared for differences in title added entries (Table 6b). At least one difference occurred in 11 out of the 14 pairs (79%) and one pair (7% of 14) had more than one difference. Three pairs (21% of 14) were consistent.

Series	# of differences	% of field differences	% of total differences
Presence/ absence	4	7.3	1
Field treatment	17	30.9	4.25
Tracing	9	16.4	2.2
Series statement	16	29.1	3.9
Numbering	4	7.3	1
MARC	5	9	1.25
Total	55	100	13.6

Table 7a: Series differences

Two series that appeared in RLIN records were absent from the OCLC records and two series that were present in OCLC records were not present in the RLIN records, which makes a total number of 4 series presence/absence differences, accounting for 7.3% of the total series differences and 1% of the total number of all differences (Table 7a). There were 17 "field treatment" differences, 30.9% of all series differences and 4.25% of all differences. There were 9 tracing differences, 16.4% of all series differences and 2.2% of all differences. Sixteen differences in series statement occured, accounting for 29.1% of all series differences and 3.9% of all differences. There were 5 MARC coding differences (9% of series differences and 1.25 % of all differences) and 4 numbering differences (7.3% of series and 1% of all differences). There were 55 total series differences among the 106 pairs, 13.6% of the total number of all differences.

# of pairs compared	38	36% of 106
# of pairs with at least 1 difference	27	71% of 38
# of pairs with >1 diff.	20	53% of 38
# of pairs consistent	11	29% of 38

Table 7b: Series consistency

From the 38 pairs (36% of 106) compared for differences in series statements, 27 pairs (71% of 38) had at least one difference (Table 7b). Twenty pairs (53% of 38) had more than one difference, and 11 pairs (29% of 38) had no differences in their series statements.

LC subject headings were analyzed for all 106 pairs. There were 254 subject headings among OCLC records and 291 subject headings among RLIN records (an average of 2.4 headings per record for OCLC and 2.7 headings per record for RLIN).

Subject headings	# of differences	% of field differences	% of total differences
Main heading	123	75	30.5
Sub-divisions	40	24.4	9.9
MARC	1	.6	.2
Total	164	100	40.7

Table 8a: Subject heading differences

A total of 164 differences were found among the subject headings of the 106 matching pairs, representing 40.7% of the total number of all differences (Table 8a). Of the total number of differences of this field, 123 were differences in the main heading, 75% of all field differences and 30.5% of the total number of all differences. A total of 40 subdivision differences accounted for 24.4% of all subject heading differences and 9.9 % of all differences. Finally, 1 MARC coding difference occurred, accounting for .6% of field differences and .2% of all differences.

# of pairs compared	106	100% of 106
# of pairs with at least 1 difference	62	58% of 106
# of pairs with >1 diff.	45	42% of 106
# of pairs consistent	44	42% of 106

Table 8b: Subject consistency

As presence of subject headings was a condition for selecting a record for this sample, all 106 pairs were compared for differences in their subject headings (Table 8b). At least one difference occurred in 62 pairs (58% of 106). Forty-five pairs (42% of 106) had more than one difference and 44 pairs (42% of 106) were consistent.

The last access points analyzed in this study were non-title added entries.

Non-title added entries	# of differences	% of field differences	% of total differences
Presence/ absence	31	67.4	7.5
Main heading	12	26.1	3.2
Sub-heading	2	4.3	.5
MARC	1	2.2	.2
Total	46	100	11.4

Table 9a: Non-title added entry differences

There were 31 presence/absence differences in non-titlte added entries, accounting for 67.4% of all differences in 7XXs and 7.5 % of all differences (Table 9a). There were also 12 main heading differences accounting for 26.1% of all differences in 7XXs and 3.2% of all differences, and 2 differences in the sub-headings which accounted for 4.3% of the field differences and .5% of all differences. Finally, there was only 1 MARC coding difference, corresponding to 2.2 % of field differences and .2 % of the total number of all differences. A total of forty-six differences were found in non-title added entries, representing 11.4% of the total number of all differences.

# of pairs compared	56	53% of 106
# of pairs with at least 1 difference	30	54% of 56
# of pairs with >1 diff.	9	16% of 56
# of pairs consistent	26	46% of 56

Table 9b: Non-title added entries consistency

Fifty-six pairs were compared for non-title added entries, accounting for 53% of 106 pairs (Table 9b). Among them 30 pairs (54% of 56) had at least one difference. Nine pairs (16% of 56) had more than one difference in their 7XX fields, and 26 pairs (46% of 56) were consistent.

A summary of MARC differences across the types of access points was also compiled. The summary showed a total of 28 differences, 6.7% of the total number of all differences (Table 10). These include 3 differences (10.7% of all MARC differences and .7 of all differences) in the 1XX field , 17 differences (60.7% of all MARC and 4.2 of all

differences) in the 245 field, 1 difference (3.6% of all MARC and .2% of all differences) in the 246/740 fields, 5 differences (17.8 of all MARC and 1.2% of all differences) in the 4XX/8XX fields, and 1 difference (3.6% of all MARC and .2% of all differences) for both 6XX and 7XX fields.

MARC	# of differences	% of MARC all differences	% of total differences
050/055/090	0	0	0
082/092	0	0	0
1XX	3	10.7	.7
130/240	0	0	0
245	17	60.7	4.2
246/740	1	3.6	.2
4XX/8XX	5	17.8	1.2
6XX	1	3.6	.2
7XX	1	3.6	.2
TOTAL	28	100	6.7

Table 10: MARC differences combined across the types of access points

In summary, twelve pairs contained differences in their LC number fields (86 % of 14 pairs compared) and two pairs (14 % of 14) were consistent in their assignment of LC numbers (Table 11a). One pair (33 % of 3 pairs compared) had differences in the assignment of DDC number and 2 pairs (67 % of 3) were consistent. Among the 106 pairs compared for differences in the choice of main entry, 7 pairs (6.6 %) contained differences and 99 pairs (93.4 %) were consistent in their choice of main entry. Sixteen pairs (20 % of 80 pairs compared) had differences in their 1XX fields, whereas 64 pairs (80 % of 80) were consistent. No differences are reported for the 130/240 fields because these fields were not present in any records. A comparison of all 106 pairs for differences

in their 245 fields showed that 45 pairs (42 %) had differences in their title statement and area of responsibilities, and 61 pairs (58 %) had no differences in this field. Twenty-seven pairs (71 % of 38 pairs compared) had differences in the series fields and 11 pairs (29 % of 38) were consistent. Sixty-two pairs (58 % of 106 pairs compared) had differences in the assignment of subject headings, and 44 pairs (42 %) had no differences. Differences occurred in 30 pairs (54 % of 56 pairs compared) in the assignment of non-title added entries, but no differences occurred in 26 pairs (46 % of 56). Finally, among all pairs compared, 101 pairs (95 %) contained differences and only 5 pairs (5 %) were consistent.

Field	# of pairs compared containing differences	% of pairs compared containing differences	# of pairs consistent	% of pairs consistent
050/055/090	12	86% of 14	2	14% of 14
082/092	1	33% of 3	2	67% of 3
main entry	7	6.6% of 106	99	93.4% of 106
1XX	16	20% of 80	64	80% of 80
130/240	0	NA	0	NA
245	45	42% of 106	61	58% of 106
246/740	11	79% of 14	3	21% of 14
4XX/8XX	27	71% of 38	11	29% of 38
6XX	62	58% of 106	44	42% of 106
Non-title 7XX	30	54% of 56	26	46% of 56
TOTAL	101	95% of 106	5	5% of 106

Table 11a: Summary of access point consistencies

A summary of the number of differences of each type of access point and its percentage of differences in relation to the total number of differences found is given in Table 11b.

	050/055/ 090	082/ 092	Main entry	1XX	130/ 240	245	246/ 740	4XX/ 8XX	6XX	7XX	Total
# of differences	28	2	7	16	0	75	11	55	164	46	404
% of Total differences	6.9	.5	1.7	3.7	0	18.8	2.7	13.6	40.7	11	100

Table 11b: Summary of access point differences

A total number of 28 differences, 6.9% out of the total number of differences, were found in the LC classification number fields (MARC fields 050, 055, 090). The comparison of the fields for DDC number (082, 092) produced a total of 2 differences, .5% of the total number of differences. Seven differences (1.7% of all differences) were found in the choice of main entry. Fields 1XX for main entry showed a total of 16 differences, 3.7% of the total number of differences. There were no 130 or 240 fields present in any of the 106 pairs studied, therefore 0 differences for these fields are reported in this table. Among the 106 pairs compared for 245 field for title statement and area of responsibility, 75 differences were found, 18.8% of the total number of all access point differences. Fields 246 and 740 produced a total number of 11 differences, 2.7% of the total number of all differences. Among the series statement fields (4XX and 8XX), 55 differences were found, 13.6% of the total number of all differences. The largest number of differences occurred in the LC subject headings fields (6XX). A total number of 164 differences were found corresponding to 40.7% of the total number of all differences. Finally, non-title added entry fields (7XX), showed 46 differences, a 11.4% of the total number of all differences.

CHAPTER V

CONCLUSIONS

The results of this study show a relatively high variability in the degree of consistency among the access points analyzed. The most surprising data in this study is the very high level of overall inconsistency among the 106 matching paired compared. Only 5% of the pairs were consistent. This could indicate that either the materials cataloged by non-LC libraries are more difficult to catalog, or, that cataloging rules and other cataloging aids are not clear enough resulting to inter-indexer inconsistency in the assignment of access points.

Considering that all of the member libraries follow AACR2 and presumably use the authority files available to them by the two databases, it is expected that access points governed AACR2, such as main entry (1XX fields), title statement and area of responsibility (245 field), title added entries (246/740 fields), and non-title added entries (7XX fields), will show a high level of inter-indexer consistency. Looking at the data we see that is true only for the choice of main entry (93% consistency) and the 1XX fields (80% consistency). Title added entries (246/740), on the other hand, show the lowest consistency (21%) among the access points governed by AACR2. It seemed that the low consistency was due to the subjectivity involved in deciding what should be traced as title added entry. Title statement and area of responsibility (245 field) also show a high level of inconsistency with a very large variability in the statement of responsibility area, where 46.58% of the 245 field differences occur. Part of this large variability is due to the fact that 11 out of 34 area of responsibility differences are differences in the completeness of

the subfield. It appears that some records have 245 field data not completely transferred to the databases. Even though in this study they were considered as differences, in reality they should probably be considered as systems errors. Still, the variability in the area of responsibility is large. One could speculate that the sources of information (title page, verso, etc.) of these materials do not present the information very clearly, or that AACR2 is unable to provide sufficient instruction in the transcription of this field.

The last type of access point analyzed in this study that is governed by cataloging rules and aids is the non-title added entries (7XX field). Forty-six percent of inconsistency among the two databases in the assignment of 7XX fields. Presence or absence of an 7XX field was the major difference (67.4% of 7XX differences) among the pairs studied for differences in the non-title added entries. One can speculate, again, that either the rules governing non-title added entries are not very clear, or that cataloger taste and judgment regarding whether to trace an 7XX field or not affects inter-indexer consistency in the assignment of non-title added entries. Looking at the level of inconsistency in the main headings of 7XX fields, one can speculate that an established form for these headings is not available through the authority files of these two databases, or that there are differences even among the two authority files. Further research would be needed to investigate these differences.

Classification numbers (LC and DDC), series statements (4XX/8XX fields), and subject heading (6XX fields) are not governed by AACR2. Therefore it is expected that the findings of this study would not show a very high level of consistency in the

assignment of the above access points, and this is indeed the case for subject heading assignment.

The largest number of differences (40.7 % of all differences) overall were found in the assignment of subject headings. Out of the 106 pairs compared 58 % were inconsistent in the assignment of subject headings. This could be due to the subjectivity involved in the decision of the topic of each document, the possibly limited level of subject knowledge of the indexers, the lack of rules in assigning specific subject headings, and the different policies each library has for the number of subject headings assigned to each item.

Looking at the data for the series fields, we see that only 29% of the pairs compared for series differences were consistent in the assignment of series. It is surprising that even though there are many differences in the series statement area, there is relatively a high consistency in the series tracing policies among the members of the two databases.

Finally, because a large number of libraries that input original cataloging did not assign any classification numbers to their items, one cannot speculate much about the findings of this study with relation to the consistency of classification number assignment. Very few libraries assigned Dewey Decimal Classification number to their books (only 2.8 % of 106 pairs were compared), and so it is not possible to make any general conclusion about the consistency in the assignment of DDC numbers. From the Library of Congress Classification numbers assigned we see a high level of inconsistency (86 % of the pairs compared for LCCN) among the two databases, although, again, the number of pairs compared was small.

Even though error studies do not look at exactly the same things as this study of differences in access points, one could hypothesize that the amount of errors in each type of access point would indicate an expected level of consistency of each type of access point. The findings of this study showed that there are, in general, few differences in MARC coding (28 differences, 6.7 % of all differences), but a large number of differences in the access points of series (55, 13.6 %), subject headings (164, 40.7 %), title statement and area of responsibility (75, 18.8 %), and non-title added entries (46, 11 %). This contradicts the findings of Intner, even though she studied cataloging errors between RLIN and OCLC, and not differences in the assignment of access points. Her study showed that 16.7 % of records contained errors in MARC tagging and only 1.2 % of records contained errors in subjects, and 23 % of records contained errors in main and added entries combined. Romero and Romero also studied cataloging errors but their study was based on entry-level cataloging. Their findings showed 5.9 % of errors in the series area, 3.4% of errors in 1XXs, 2.25 % of errors in added entries, 7.9 % of errors in subject headings and 13.3 % of errors in MARC format of variable fields and 27.2 % of errors in MARC format of fixed fields. Romero and Romero's findings of errors in MARC tagging of fixed fields could explain the large difference in the findings of MARC tagging areas between Intner and this study which could be due to differences in fixed fields which were not studied in this present study.

Comparing the findings of the above two error studies with the present inter-indexer consistency study we see that the number of errors found in each type of access

point are not indicative of the level of inter-indexer consistency of the respective type of access point.

Because few comparisons were able to be made regarding classification number assignment, title added entry assignment and no comparisons made regarding uniform title assignment, future research using a much larger sample would be desirable. Another area that needs to be studied is inter-indexer consistency in the assignment of access points for non-book materials, where different conditions and cataloging factors apply. Very few studies have focused in the comparison of the cataloging of non-book materials.

It would be also very helpful if studies similar to inter-indexer consistency were conducted to analyze intra-indexer consistency. It is very important to see whether the same cataloger will assign the same access points if they catalog the same item at two different times. Such a study would place inter-indexer consistency studies in perspective and will also give some insight to information indexing studies in general.

REFERENCES

Blair, David C. 1986. Indeterminacy in the subject access to documents. Information Processing and Management 22 (2) : 229-241.

Chan, Lois Mai. 1989. Inter-indexer consistency in subject cataloging. Information Technology and Libraries 8 (4) : 349-358.

Chen, H , A K. Danowitz, K J. Lynch, S E. Goodman, and W.K. McHenry. 1994. Explaining and alleviating information management indeterminism: A knowledge-based framework. Information Processing and Management 30 (4) : 557-577.

Furnas, George W , Louis M. Gomez, Thomas K. Landauer, and Susan T. Dumais. 1982. Statistical semantics: How can a computer use what people name things to guess what things mean when they name things? In Human Factors in Computer Systems, CHI '82 Conference Proceedings, March 15-17, 1982, Gaithersburg, Maryland. New York: Association for Computing Machinery, 251-253.

Herr, Joanne J. 1970. Use of data-base access for interindexer communication and for indexer training. In Proceedings of the American Society for Information Science: 33rd Annual Meeting, Philadelphia, October 11-15, 1970. Washington, D.C.: ASIS, 163-166.

Intner, Shelia S. 1989. Much ado about nothing: OCLC and RLIN cataloging quality. Library Journal 114 (February 1) : 38-40.

Intner, Sheila S. 1989. Quality in bibliographic databases: An analysis of memeber-contributed cataloging in OCLC and RLIN. Advances in Library Administation and Organization 8 : 1-24.

Leonard, L. E. 1977. Inter-indexer consistency studies, 1954-1975: A review of the literature and summary of the study results. University of Illinois: Graduate School of Library Science.

MacMillan, Judith T., and Isaac D. Welt. 1961. A study of indexing procedures in a limited area of the medical sciences. American Documentation 12 : 27-31.

Markey, Karen. 1984. Interindexer consistency tests: A literature review and report of a test of consistency in indexing of visual materials. Library & Information Science Research 6 : 155-177.

O'Neill, E.T., and A. Rao. 1979. Research report on subject heading patterns in OCLC monographic records. Columbus, Ohio: OCLC, Inc.

Romero, Lisa, and Nancy Romero. 1992. Original cataloging in a decemtralized environment: An identification and explanation of errors. Cataloging & Classification Quarterly 15 (4) : 47-65.

Romero, Lisa. 1984. Original cataloging errors: A comparison of errors found in entry-level cataloging with errors found in OCLC and RLIN." Technical Services Quarterly 12 (2) : 13-27.

Slamecka, Vladimir, and Joan E. Jacoby. 1965. Consistency of human indexing. In The Coming Age of Information Technology. Bethesda: Documentation Inc., 32-56.

Tarr, Daniel, and Harold Borko. 1974. Factors influencing inter-indexing consistency. In Information utilities: Proceedings of the 37th Annual Meeting, Atlanta, Georgia, October 13-17, 1974. Washington, D.C.: American Society for Information Science, 50-55.

Tell, Bjorn V., and AB Atonenergi. 1969. Document representation and indexer consistency: A study of indexing from title, abstract, and full text using UDC and keywords. In Proceedings of the American Society for Information Science: 32nd Annual Meeting, San Francisco, California, October 1-4, 1969. Westport, Conn.: Greenwood Publishing Co., 285-292.

Zunde, Pranas, and Margaret E. Dexter. 1969. Factors affecting indexing performance. In Proceedings of the American Society for Information Science: 32nd Annual Meeting, San Francisco, California, October 1-4, 1969. Westport, Conn.: Greenwood Publishing Co., 313-322.

APPENDIX

Worksheet

090/050/055

Presence: OCLC ____ RLIN ____
Class difference ____
Class number difference ____
Cutter # ____ Date _____ diff.

092/082

Presence: OCLC ____ RLIN ____
Class difference ____
Cutter # ____ Date diff. ____

1XX

Presence: OCLC ____ RLIN ____
Heading difference ____
MARC diff.: Field ____
Subf. ____

130, 240

Title statement difference ____
MARC: Field diff. ____
Subf. diff. ____

245

Title statement difference ____
Subtitle ___ Responsibility ____
Main entry difference ____
MARC diff. ____

246, 740

Field difference ____
Title statement difference ____
# of title added entries:
OCLC ____ RLIN ____
Presence/absence ____
MARC coding diff. ____

4XX, 8XX

Field treatment difference ____
Tracing difference ____
Series statement difference ____
MARC coding diff. ____
Presence: OCLC ___ RLIN ___

6XX

Main heading difference ____
Subheading difference ____
# of 6XXs: OCLC ___ RLIN ____
MARC coding diff. ____

7XX

Field difference ____
Main heading difference ____
Subheading difference ____
# of 7XXs: OCLC ___ RLIN ___
Presence/absence ____
MARC coding diff. ____